Journal of Computer Science

Original Research Paper

A Hybrid Feature Extraction Method for Accuracy Improvement in "Aksara Lontara" Translation

[1#]Intan Sari Areni, [2#]Asyraful Insan Asry and [3#]Indrabayu

[1,2]Department of Electrical Engineering, Universitas Hasanuddin, Indonesia
[3]Department of Informatics, Universitas Hasanuddin, Indonesia
[#]Research Group of Artificial Intelligence and Multimedia Processing (AIMP)

Article history
Received: 19-04-2017
Revised: 15-07-2017
Accepted: 26-07-2017

Corresponding Author:
Intan Sari Areni,
Electrical Study Program,
Department of Electrical Engineering, Universitas Hasanuddin, Makassar, Indonesia
Email: intan@unhas.ac.id

Abstract: An Optical Character Recognition (OCR) of "Aksara Lontara" has been constructed using a novel combination of feature extraction methods in this study. The ancient font of "Lontara" is then translated into Bahasa Indonesia to help non-native language to learn this language. Two powerful extraction feature methods, i.e., Modified Direction Feature (MDF) and Fourier Descriptor (FD) are stages combined to deal with two dominant phases of the Lontara font. The classification process is conducted using Support Vector Machine (SVM) as a fast and straightforward learning method deal with 23 fonts in image containing of 150×120 pixels. In this research, 50 verbs were used for training and 30 verbs for validating the system. The results show that system can reach 96% accuracy using this hybrid in extraction feature with kernel variable of $C = 3$ and $\sigma = 8$.

Keywords: Lontara script, Modified Direction Feature, Fourier Descriptor, Support Vector

Introduction

One of a great and most implemented field in pattern recognitions is Optical Character Recognition (OCR) which can transform images into a script for further processing by computers (Mori *et al*., 1992). Nowadays, this OCR technique is also exploited for non-Latin fonts such as Japanese kanji (Budiwati *et al*., 2011) and Chinese (Hao *et al*., 2011). Due to uniqueness in font pattern of each language, a researcher needs to find and explore the best methods from preprocessing until recognition stages.

"Aksara Lontara" is one of the ageless legacies of Indonesia, particularly in Bugis and Makassar tribes. It is written in lontara leaves and preserves until now. "Lagaligo" is one of the famous books written in lontara, fills with poetry and philosophy of two biggest kingdoms at the time i.e., Buginese and Makassarnese (Mattulada, 1991). An effort for digitalization and embedding this literacy into compact and straightforward form is vital for the new generation. This portable system will be useful not only for scholars but also for tourism purposes related to Aksara Lontara.

Several researchers have engaged in developing an OCR for aksara lontara through assimilating image processing techniques and artificial intelligence. It has been completed for specific characters and classification purposes only using Backpropagation Neural Network (Alwi and Wardoyo, 2010). Moreover, a web base of aksara lontara for translation was adapted using parsing tree and binary search methods (Nangi *et al*., 2013). Modified Direction Feature (MDF) was proposed for in aksara Lampung recognition which is limited to stand-alone primary characters and without punctuation (Prarian *et al*., 2013). MDF is also favorable for aksara Bali with distinct different parameter transition and normalization condition in feature extraction stage. In that paper, Learning Vector Quantization is used for classification step (Tjokorda *et al*., 2009). Another method, Fourier Descriptor (FD), was applied in detection of handwritten Aksara Batak Toba using Multilayer Perceptron algorithm neural network (Sihombing, 2013), which has been multiplied somewhat for classification stage in artificial intelligence application (Indrabayu *et al*., 2013).

Previous researches related to aksara recognition just focus on recognize one character without punctuation and MDF method was well performed on that. In this study, the authors proposed to recognize two characters of aksara Lontara that have special rules in reading and

Science Publications

writing. In this study, authors proposed a hybrid feature extraction, MDF and FD methods, for a translation system of "Aksara Lontara" based on image script. This is a uniqueness of this article. FD method can detect pixels of shape character (Rajput and Horakeri., 2011) especially for punctuation shapes in Lontara word.

The translation system is built for Indonesia tourism especially in Makassar.

The rest of this paper is organized as follows; Section 2 describes methodology include discussion of the hybrid feature extraction implementation. Section 3 explained results and discussion of the case study. Section 4 describes our conclusion and some future works.

Methodology

Words in Lontara formed by two or more Lontara symbols which have special rules in reading and writing. This study focuses on the translation of the verb which is composed of two symbols of lontara. Figure 1 shows the character of Lontara script.

Image of Lontara verb is used as data input with a size of 150×120 in JPG format. Font type is BugisA.ttf with a size of 14 pt. The classification stage consists of two processes, i.e., training and testing process. The number of the used words for training and testing is 50 and 30, respectively. The training process is applied to get the best parameters of the feature values which is obtained by the extraction process of MDF and FD methods. When the training process has been carried out, the best outcome parameters is saved and observed for the further classification of testing data. The result of the testing process is associated with a database for the translation process. The proposed system design is shown in Fig. 2.

Preprocessing

Preprocessing consists of grayscaling and binarization. Grayscaling is the process of converting the pixel values of RGB images to grayscale pixel value based on the following equation:

$$Gray = w_{R.}R + w_{G.}G + w_{B.}B \qquad (1)$$

where, w_R, w_G and w_B are the weight value of Red, Green and Blue color, respectively. According to National Television System Committee (NTSC), w_R is 0.299, w_G is 0.587 and w_B are 0.114. Figure 3 shows a sample of the RGB values for the word "Akka" at the point marked.

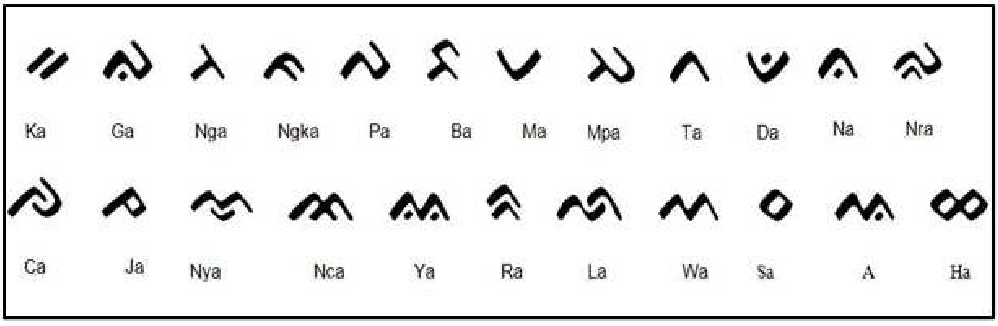


Fig. 1. Character of Lontara script

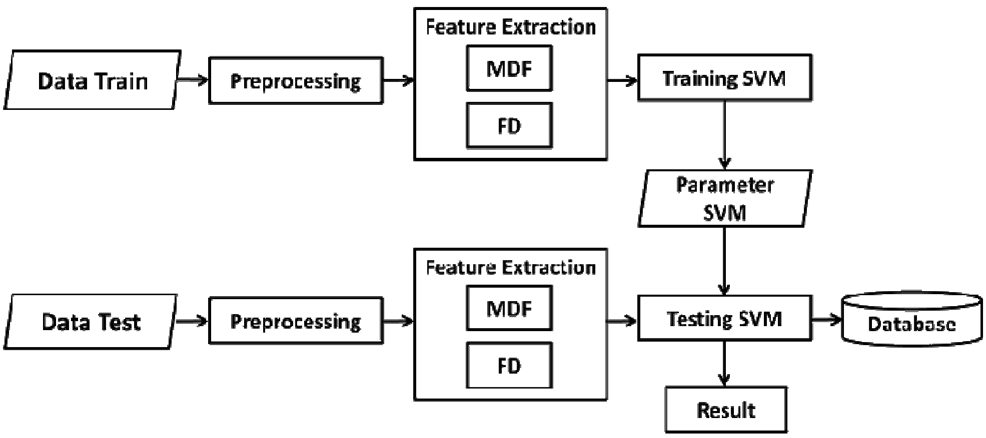


Fig. 2. Proposed system design

DOI: 10.3844/jcssp.2017.393.399

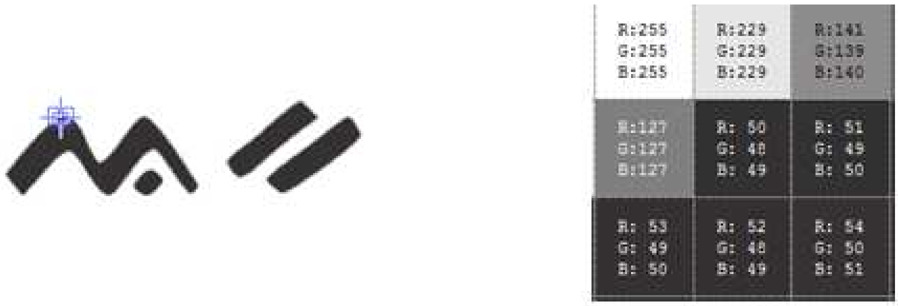


Fig. 3. The example of RGB values for "Akka"

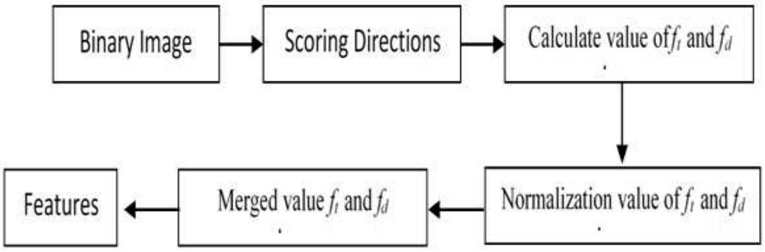


Fig. 4. Flow diagram of MDF

Table 1. The value of label directions (d_n)

Direction	Label	Shape
Vertical	2	|
Right Diagonal	3	/
Horizontal	4	—
Left Diagonal	5	\

MDF method is combination feature method of FD (Fourier Descriptor) and FT (Fourier Transform). In FD, the feature value is determined by the direction value (d_n) as shown in Table 1.

After the gray scaling process, each pixel value is converted into pixel values range from 0-255. From this point, the grayscale value $G(x,y)$ of each pixel is converted into a binary value $B(x,y)$, which is called binarization process, with a threshold value (T) of 210 and based on the following equation:

$$B(x,y)=\begin{cases}1 if\ G(x,y)\leq T\\ 0 if\ G(x,y)\geq T\end{cases} \tag{2}$$

Hybrid Feature Extraction

Modified Direction Feature (MDF) (Liu and Blumenstein, 2004) and Fourier Descriptor (FD) (Kadir & Susanto, 2013) methods are used in feature extraction stage. At the MDF, there are two feature values (N_f), namely direction feature value (f_d) and the transition function value (f_t) where a flow diagram of MDF is shown in Fig. 4.

Furthermore, the value of f_t is calculated by the following equation:

$$f_t = 1-(p / w) \tag{3}$$

where, p is pixel position and w is image width. While the value of f_d is determined by the direction (d_n) with the following equation:

$$f_d = 0.1\ d_n \tag{4}$$

The d_n value is derived from Table 1 based on the direction value of object-forming pixels. Figure 5 shows the d_n value example of the word "Akka". While the f_t and f_d values for the word of "Akka" is shown in Fig. 6.

The value of f_t and f_d is normalized in each direction with a normalization matrix size (N_m) of 5×3 and those values are combined from all sides so that the resulting vector length is 120 (= $N_m \times N_f \times N_d$).

After the features extraction results obtained by the MDF, the following process is also conducted using FD. The features extraction with FD is determined by the index value of the object contour pixels and converted to a form of vector value: $x + jy$. The results of the vector values $s(t)$ are transformed by the Fast Fourier Transform (FFT) as in the following equation:

$$Un=\frac{1}{N}\sum_{t=0}^{N}s(t).\exp(\frac{-j2\pi t}{N}), n=0,1,2,...,N-1 \tag{5}$$

where, N is the number of image pixels, U_n is the Fourier descriptors which consist of the object pixel values with $n = 0, 1, 2, .., N-1$. The FD stage is shown in Fig. 7.

The combination of features extraction concept is shown in Fig. 8. The MDF method can detect every pixel in a gray area, but not in a small contour (red punctuation). The deficiencies in the MDF can be overcome by the FD method. Thus, the combination of both methods will improve the performance of the proposed translator system.

0	0	0	0	0	0	0	2	0
0	0	0	0	3	4	4	2	4
0	0	0	3	3	4	4	2	4
0	0	3	3	3	4	4	0	4
0	3	3	3	3	0	0	0	4

Fig. 5. The label directions object value of the word "Akka"

(a)

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.97	0
0	0	0	0	0.96	0.95	0.95	0.97	0.95
0	0	0	0.96	0.96	0.95	0.95	0.97	0.95
0	0	0.96	0.96	0.96	0.95	0.95	0	0.95
0	0.96	0.96	0.96	0	0	0	0	0.95

(b)

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.2	0
0	0	0	0	0.3	0.4	0.4	0.2	0.4
0	0	0	0.3	0.3	0.4	0.4	0.2	0.4
0	0	0.3	0.3	0.3	0.4	0.4	0	0.4
0	0.3	0.3	0.3	0	0	0	0	0.4

Fig. 6. Feature values: (a) Transition feature value (f_t) and (b) Direction feature value (f_d)

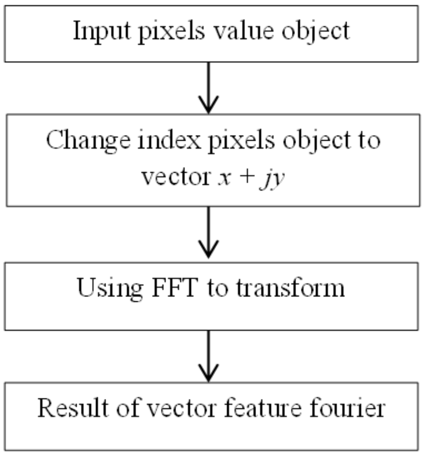


Fig. 7. Fourier descriptor stages

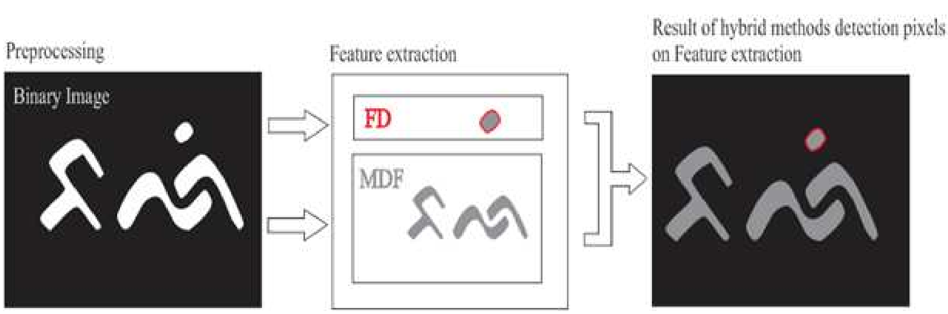


Fig. 8. The combination of feature extraction concept

Classification

The resulting feature values on all training data will be combined and used as an input data in the training process with Support Vector Machine (SVM). SVM method aims to lock the best hyperplane to get the path separator from two different classes (Cheriet *et al*., 2007). This study uses SVM multi-classes to classify each feature values with the estimation method of one to one so that the separation function of the two classes can be determined. There are two parameters must be considered in SVM training with the *Radial Basis Function* (RBF) kernel, i.e., C and γ. The C parameter controls the influence of each individual support vector and the γ parameter defines how far the influence of a single training example reaches.

Furthermore, the training data is expressed as (x_i, y_j), with $i = 1, 2, 3, \ldots, M$, where M is the number of data and denoted by the model of $x_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iq}\}^T$ that is an attribute to i-th training data. Class labeling is denoted by $y_j \in \{-1,+1\}$, where $j = 1, 2, 3, \ldots, P$. A class of data samples separated by hyperplane is expressed by the following equation:

$$f(x) = w\,x + b \tag{6}$$

where, w is weighting vector and b is a constant value. where, $\gamma = 1/2\sigma^2$ with σ is a free parameter.

The performance of translation system is expressed with an accuracy level (A_L) based on the following equation:

$$A_L = (N_r / N_t) \times 100\% \tag{7}$$

where, N_r is the amount of correct data and N_t is the amount of all data.

Results and Discussions

In the testing process, the used testing data is 30 by varying parameters C and σ in kernel RBF. Figure 9 shows the accuracy of the MDF and MDF-FD for any difference parameter.

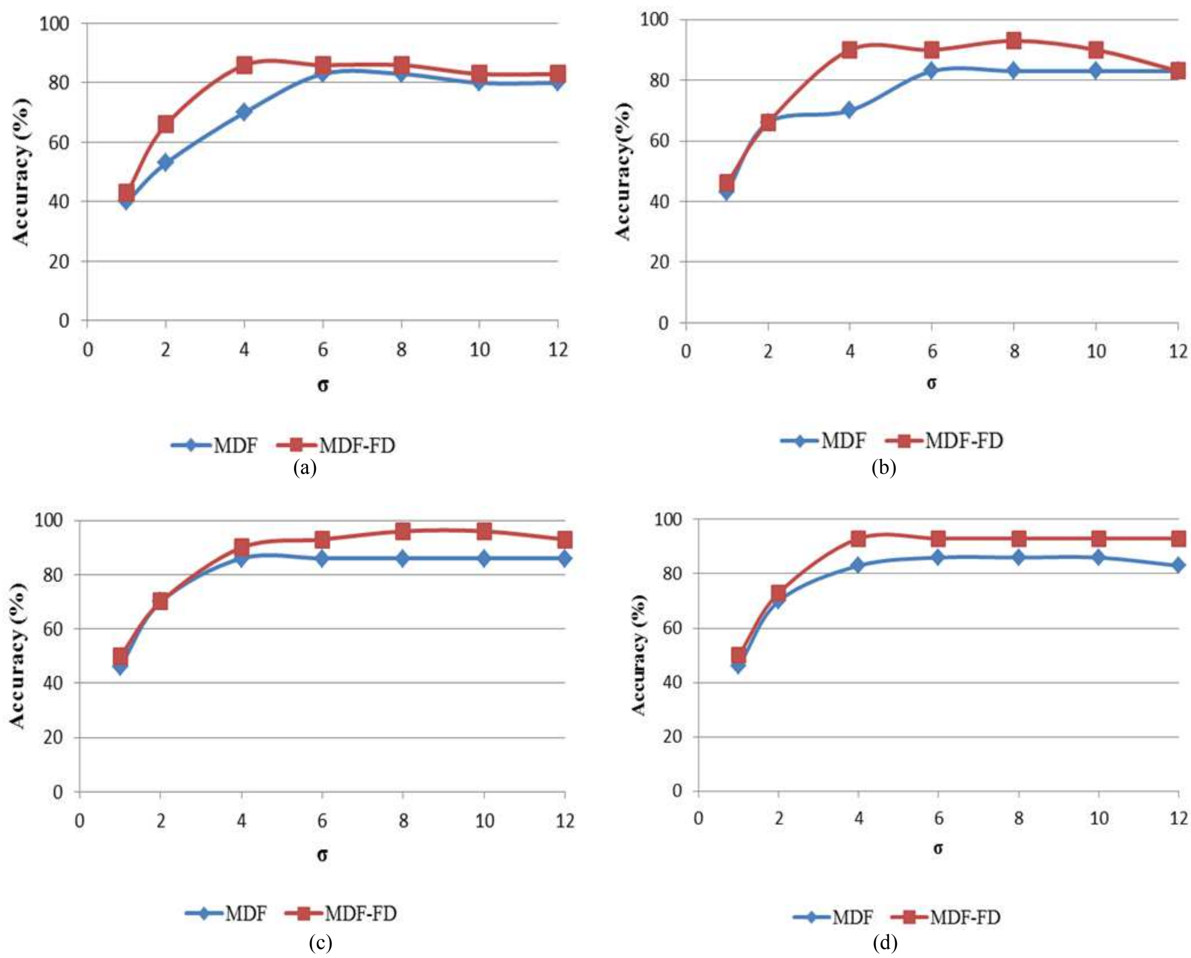
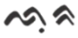


Fig. 9. The comparison accuracy of the translation system; (a) $C = 1$; (b) $C = 2$; (c) $C = 3$; (d) $C = 4$

Table 2. The results of the testing data

No	Word Aksara	Spelling	Meaning	MDF	MDF-FD
1		*Jamma*	Adu	Adu	Adu
2		*Bissa*	Basuh	Bicara	Basuh
3		*Akka*	Angkat	Angkat	Angkat
4		*Tunu*	Bakar	Bakar	Bakar
5		*Dince*	Tancap	Tancap	Tancap
6		*Giling*	Balik	Balik	Balik
7		*Mala*	Ambil	Ambil	Ambil
8		*Lurang*	Angkut	Angkut	Angkut
9		*Tojang*	Ayun	Ayun	Ayun
10		*Era*	Ajak	Ajak	Ajak
11		*Walek*	Balas	Balas	Balas
12		*Jallo*	Amuk	Amuk	Amuk
13		*Tahang*	Tahan	Tahan	Tahan
14		*Bici*	Bisik	Bisik	Bisik
15		*Bali*	Bantu	Makan	Bantu
16		*Nyikka*	Ikat	Makan	Ikat
17		*Pau*	Bicara	Bicara	Bicara
18		*Ekbu*	Buat	Buat	Buat
19		*Bampa*	Pukul	Pukul	Pukul
20		*Manre*	Makan	Makan	Makan
21		*Gerrak*	Bentak	Bentak	Bentak
22		*Sungke*	Buka	Buka	Buka
23		*Boya*	Cari	Cari	Cari
24		*Nange*	Renang	Renang	Renang
25		*Sare*	Beri	Beri	Beri
26		*Tawa*	Bagi	Bagi	Bagi
27		*Basa*	Bicara	Bicara	Bicara
28		*Anang*	Anyam	Angkat	Angkat
29		*Puwe*	Belah	Belah	Belah
30		*Tettong*	Berdiri	Berdiri	Berdiri

In Fig. 9, the best accuracy using MDF is 86% when C = 3 and σ = 8, which is improved by using the combination of feature extraction MDF-FD and scaling up to 96% with the same parameter. Table 2 shows the results of each data testing.

Misclassification with MDF feature extraction occurs in the data that has punctuation on top due to the normalization value of the f_d and f_t will tend to ignore small fractal value. This issue will lead to mistranslation

Conclusion

The process of feature extraction with a combination of MDF-FD system has been applied to the translation system on Lontara script with image-based data input. The total amount of data process is 80 comprises of 50 training data and 30 testing data. The results showed that the performance of the translation system increased 10% (from 86 to 96%) with a hybrid features extraction when C = 3 and σ = 8. This is due to FD method can recognize the small shape in Lontara script that is not detected by MDF. Hence, the combination of both methods can improve the system performance.

In the future research, this system will incorporated more data, not only for verbs but also nouns. Words meaning will also seek into single and two connected words. Moreover, this system will be verified by other methods, such as Wavelet and Shearlet transforms.

For tourism purposes, this study will be improved by building an Android-based application to get a portable and user-friendly translation tools.

Acknowledgement

This work was supported by WCU Grant of Universitas Hasanuddin.

Author's Contribution

All authors equally contributed to this work.

Ethics

This article is the original contribution of the authors and is not published elsewhere. There is no ethical issue involved in this article.

References

Alwi, A. and R. Wardoyo, 2010. Character recognition of lontara bugis-makassar using backpropagation neural network method. Unpublished Theses for Graduate School of Gadjah Mada University, Yogyakarta, Indonesian.

Budiwati, S.D., J. Haryatno and E.M. Dharma, 2011. Japanese character (Kana) pattern recognition application using neural network. Proceedings of the International Conference on Electrical Engineering and Informatics, Jul. 17-19, IEEE Xplore Press, Bandung, Indonesia, pp: 1-6. DOI: 10.1109/ICEEI.2011.6021648

Cheriet, M., N. Kharma, C.L. Liu and C. Suen, 2007. Character Recognition Systems: A Guide for Students and Practitioners. 1st Edn., John Wiley and Son, Hoboken, ISBN-10: 0470176520, pp: 351.

Hao, H.W., C.X. Mu, X.C. Yin and Z.B. Wang, 2011. Handwritten Chinese character identification with bagged one-class support vector machines. Proceedings of the International Joint Conference on Neural Networks, Jul. 31-Aug. 5, IEEE Xplore Press, USA., pp: 159-164. DOI: 10.1109/IJCNN.2011.6033215

Indrabayu, A., N. Harun, M.S. Pallu and A. Achmad, 2013. Statistic approach versus artificial intelligence for rainfall prediction based on data series. Int. J. Eng. Technol., 5: 1962-1969.

Kadir, A. and A. Susanto, 2013. Theory and Image Processing Applications. 1st Edn., Andi Publisher, Yogyakarta, ISBN-13: 9789792934304.

Liu, X.Y. and M. Blumenstein, 2004. Experimental analysis of the modified direction feature for cursive character recognition. Proceedings of the International Joint Conference on Neural Networks, Oct. 26-29, IEEE Xplore Press, Japan, pp: 353-358. DOI: 10.1109/IWFHR.2004.40

Mattulada, 1991. Human and culture of Bugis-Makassar in Sulawesi and Kaili. Indonesia Soc. Cult. Anthropol. J.

Mori, S., C.Y. Suen and K. Yamamoto, 1992. Historical review of OCR research and development. Proc. IEEE, 80: 1029-1058. DOI: 10.1109/5.156468

Nangi, J., N. Harun and Syafaruddin, 2013. Dictionary translator application system Indonesian-Aksara Bugis Lontara method using binary search and parsing tree. Faculty of Engineering, Universitas Hasanuddin, Makassar.

Prarian, C., I.T. Iwan and R. Magdalena, 2013. Character recognition of Aksara lampung analysis using Modified Direction Feature (MDF) and hidden Markov Model. Faculty of Electrical Engineering, Telkom Unversity. Bandung.

Rajput, G.G. and R. Horakeri, 2011. Shape descriptors based handwritten character recognition engine with application to Kannada characters. Proceedings of the International Conference on Computer and Communication Technology, Sept. 15-17, IEEE Xplore Press, Allahabad, India, pp: 135-141. DOI: 10.1109/ICCCT.2011.6075175

Sihombing, J.R., 2013. Handwritten Batak toba alphabet recognition using multilayer perceptron algorithm neural network. Technical Engineering, Kristen Maranatha University.

Tjokorda, A., I.G. Hermanto and N. Retno, 2009. Character recognition of bali using modified direction feature method and learning vector quantization. Proceedings of the National Conference Systems and Informatics, (CSI' 09), Bali, pp: 7-12.